# Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks

Tschandl, Philipp ; Rosendahl, Cliff ; Akay, Bengu Nisa ; Argenziano, Giuseppe ; Blum, Andreas ; Braun, Ralph P ; Cabo, Horacio ; Gourhant, Jean-Yves ; Kreusch, Jürgen ; Lallas, Aimilios ; Lapins, Jan ; Marghoob, Ashfaq ; Menzies, Scott ; Neuber, Nina Maria ; Paoli, John ; Rabinovitz, Harold S ; Rinner, Christoph ; Scope, Alon ; Soyer, H Peter ; Sinz, Christoph ; Thomas, Luc ; Zalaudek, Iris ; Kittler, Harald

Abstract: Importance: Convolutional neural networks (CNNs) achieve expert-level accuracy in the diagnosis of pigmented melanocytic lesions. However, the most common types of skin cancer are nonpigmented and nonmelanocytic, and are more difficult to diagnose. Objective: To compare the accuracy of a CNN-based classifier with that of physicians with different levels of experience. Design, Setting, and Participants: A CNN-based classification model was trained on 7895 dermoscopic and 5829 close-up images of lesions excised at a primary skin cancer clinic between January 1, 2008, and July 13, 2017, for a combined evaluation of both imaging methods. The combined CNN (cCNN) was tested on a set of 2072 unknown cases and compared with results from 95 human raters who were medical personnel, including 62 board-certified dermatologists, with different experience in dermoscopy. Main Outcomes and Measures: The proportions of correct specific diagnoses and the accuracy to differentiate between benign and malignant lesions measured as an area under the receiver operating characteristic curve served as main outcome measures. Results: Among 95 human raters (51.6% female; mean age, 43.4 years; 95% CI, 41.0-45.7 years), the participants were divided into 3 groups (according to years of experience with dermoscopy): beginner raters (<3 years), intermediate raters (3-10 years), or expert raters (>10 years). The area under the receiver operating characteristic curve of the trained cCNN was higher than human ratings (0.742; 95% CI, 0.729-0.755 vs 0.695; 95% CI, 0.676-0.713; $P < .001$). The specificity was fixed at the mean level of human raters (51.3%), and therefore the sensitivity of the cCNN (80.5%; 95% CI, 79.0%-82.1%) was higher than that of human raters (77.6%; 95% CI, 74.7%-80.5%). The cCNN achieved a higher percentage of correct specific diagnoses compared with human raters (37.6%; 95% CI, 36.6%-38.4% vs 33.5%; 95% CI, 31.5%-35.6%; $P = .001$) but not compared with experts (37.3%; 95% CI, 35.7%-38.8% vs 40.0%; 95% CI, 37.0%-43.0%; $P = .18$). Conclusions and Relevance: Neural networks are able to classify dermoscopic and close-up images of nonpigmented lesions as accurately as human experts in an experimental setting.

DOI: https://doi.org/10.1001/jamadermatol.2018.4378

Tschandl, Philipp; Rosendahl, Cliff; Akay, Bengu Nisa; Argenziano, Giuseppe; Blum, Andreas; Braun, Ralph P; Cabo, Horacio; Gourhant, Jean-Yves; Kreusch, Jürgen; Lallas, Aimilios; Lapins, Jan; Marghoob, Ashfaq; Menzies, Scott; Neuber, Nina Maria; Paoli, John; Rabinovitz, Harold S; Rinner, Christoph; Scope, Alon; Soyer, H Peter; Sinz, Christoph; Thomas, Luc; Zalaudek, Iris; Kittler, Harald (2019). Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. JAMA Dermatology, 155(1):58.
DOI: https://doi.org/10.1001/jamadermatol.2018.4378

JAMA Dermatology | Original Investigation

# Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks

Philipp Tschandl, MD, PhD; Cliff Rosendahl, PhD; Bengu Nisa Akay, MD; Giuseppe Argenziano, MD, PhD; Andreas Blum, MD; Ralph P. Braun, MD, PhD; Horacio Cabo, MD, PhD; Jean-Yves Gourhant, MD; Jürgen Kreusch, MD, PhD; Aimilios Lallas, MD; Jan Lapins, MD, PhD; Ashfaq Marghoob, MD; Scott Menzies, MBBS, PhD; Nina Maria Neuber, MD; John Paoli, MD, PhD; Harold S. Rabinovitz, MD; Christoph Rinner, PhD; Alon Scope, MD; H. Peter Soyer, MD; Christoph Sinz, MD; Luc Thomas, MD, PhD; Iris Zalaudek, MD; Harald Kittler, MD

**IMPORTANCE** Convolutional neural networks (CNNs) achieve expert-level accuracy in the diagnosis of pigmented melanocytic lesions. However, the most common types of skin cancer are nonpigmented and nonmelanocytic, and are more difficult to diagnose.

**OBJECTIVE** To compare the accuracy of a CNN-based classifier with that of physicians with different levels of experience.

**DESIGN, SETTING, AND PARTICIPANTS** A CNN-based classification model was trained on 7895 dermoscopic and 5829 close-up images of lesions excised at a primary skin cancer clinic between January 1, 2008, and July 13, 2017, for a combined evaluation of both imaging methods. The combined CNN (cCNN) was tested on a set of 2072 unknown cases and compared with results from 95 human raters who were medical personnel, including 62 board-certified dermatologists, with different experience in dermoscopy.

**MAIN OUTCOMES AND MEASURES** The proportions of correct specific diagnoses and the accuracy to differentiate between benign and malignant lesions measured as an area under the receiver operating characteristic curve served as main outcome measures.

**RESULTS** Among 95 human raters (51.6% female; mean age, 43.4 years; 95% CI, 41.0-45.7 years), the participants were divided into 3 groups (according to years of experience with dermoscopy): beginner raters (<3 years), intermediate raters (3-10 years), or expert raters (>10 years). The area under the receiver operating characteristic curve of the trained cCNN was higher than human ratings (0.742; 95% CI, 0.729-0.755 vs 0.695; 95% CI, 0.676-0.713; $P < .001$). The specificity was fixed at the mean level of human raters (51.3%), and therefore the sensitivity of the cCNN (80.5%; 95% CI, 79.0%-82.1%) was higher than that of human raters (77.6%; 95% CI, 74.7%-80.5%). The cCNN achieved a higher percentage of correct specific diagnoses compared with human raters (37.6%; 95% CI, 36.6%-38.4% vs 33.5%; 95% CI, 31.5%-35.6%; $P = .001$) but not compared with experts (37.3%; 95% CI, 35.7%-38.8% vs 40.0%; 95% CI, 37.0%-43.0%; $P = .18$).

**CONCLUSIONS AND RELEVANCE** Neural networks are able to classify dermoscopic and close-up images of nonpigmented lesions as accurately as human experts in an experimental setting.

*JAMA Dermatol*. doi:10.1001/jamadermatol.2018.4378
Published online November 28, 2018.

Supplemental content

**Author Affiliations:** Author affiliations are listed at the end of this article.

**Corresponding Author:** Philipp Tschandl, MD, PhD, ViDIR Group, Department of Dermatology, Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria (philipp.tschandl@meduniwien.ac.at).

In comparison with inspection with the unaided eye, dermoscopy (dermatoscopy[1]) improves the accuracy of the diagnosis of pigmented skin lesions.[2] The improvement of dermoscopy is most evident for small and inconspicuous melanomas[3] and for pigmented basal cell carcinoma.[4] Because dermoscopic criteria are more specific and the number of differential diagnoses is significantly lower, pigmented skin lesions are easier to diagnose than nonpigmented lesions. The most common types of skin cancers, however, are usually nonpigmented. A previous study showed that dermoscopy also improves the accuracy of the diagnosis of nonpigmented lesions, although the improvement was less pronounced than for pigmented lesions.[5] The proportion of correct diagnoses by expert raters increased from 41.3% with the unaided eye to 52.7% with dermoscopy. The improvement of nonexperts was less pronounced.

Artificial neural networks have been used for automated classification of skin lesions for many years[6-8] and have also been tested prospectively.[9] In comparison with the neural networks that were used before 2012,[7,10] current convolutional neural networks (CNNs) consist of convolutional filters, which are able to detect low-level structures such as colors, contrasts, and edges. These filters allow the CNNs to be trained "end-to-end," which means that they need only raw image data as input without any preprocessing, such as segmentation or handcrafted feature extraction. Unlike classical artificial neural networks, which have neurons that are fully connected to all neurons at the next layer, CNNs use reusable filters that dramatically simplify the network connections, which makes them more suitable for image classification tasks. After Krizhevsky et al[11] demonstrated in 2012 that CNNs can be trained on 1.2 million images[12] to classify 1000 categories with high accuracy, CNNs were increasingly applied to medical images. Convolutional neural networks have shown expert-level performance in the classification of skin diseases on clinical images[13,14] and in the classification of pigmented lesions on dermoscopic images.[15-18] Other research groups have combined clinical and dermoscopic image analysis[19,20] or integrated patient metadata[21,22] to improve the performance of CNNs on pigmented lesions.

The differentiation of melanoma from benign pigmented lesions is a simple binary classification problem. Performing a differential diagnosis of nonpigmented lesions, however, is a more complex, multiclass classification problem[14,15,19] that includes a range of different diagnostic categories, such as benign and malignant neoplasms, cysts, and inflammatory diseases. Automated classifiers have been applied successfully to clinical images to diagnose nonpigmented skin cancer,[23] and CNNs trained on clinical images have recently shown promising results when compared with physicians' diagnoses.[13,14] Because the performance of CNNs on dermoscopic images of nonpigmented skin lesions is still unknown, we trained and tested a CNN on a large set of nonpigmented lesions with a wide range of diagnoses and compared the results with the accuracy of human raters with different levels of experience, including 62 board-certified dermatologists.

### Key Points

**Question** Can a neural network classify nonpigmented skin lesions as accurately as human experts?

**Findings** In this study, a combined convolutional neural network that received dermoscopic and close-up images as inputs achieved a diagnostic accuracy on par with human experts and outperformed beginner raters and intermediate raters.

**Meaning** In an experimental setting, a combined convolutional neural network can outperform human raters, but the lack of accuracy for rare diseases limits its application in clinical practice.

## Methods

### Image Data Sets

The 7895 dermoscopic and 5829 close-up images of the training set originated from a consecutive sample of lesions photographed and excised by one of us (C.R.) at a primary skin cancer clinic in Queensland, Australia, between January 1, 2008, and July 13, 2017. The 340 dermoscopic and 635 close-up images of the validation set were extracted from educational slides and are part of a convenience sample of lesions photographed and excised in the practice of one of us (H.S.R). Dermoscopic images and clinical close-up images were taken with different cameras and dermatoscopes at different resolutions in polarizing or nonpolarizing mode. Pathologic diagnoses were merged to correspond to the categories used in the study by Sinz et al.[5] Use of the images is based on ethics review board protocols EK 1081/2015 (Medical University of Vienna) and 2015000162 (University of Queensland). Rater data from the survey were collected in a deidentified fashion; therefore written consent was not required by the ethics review board of the Medical University of Vienna.

We included cases that fulfilled the following criteria: (1) lack of pigment, (2) availability of at least 1 clinical close-up image or 1 dermatoscopic image, and (3) availability of an unequivocal histopathologic report. We excluded mucosal cases, cases with missing or equivocal histopathologic reports, cases with low image quality, and cases of diagnostic categories with fewer than 10 examples in the training set. All images were reviewed manually by 2 of us (H.K. and C.S.) and were included only if they conformed to the imaging standards published previously.[24] Close-up images were taken with a spacer attached to a digital single-lens reflex camera removing all incident light and standardizing distance and field of view. The diagnostic categories used for training were the following: actinic keratoses and intraepithelial carcinoma (also known as Bowen disease), basal cell carcinoma (all subtypes), benign keratosis-like lesions (including solar lentigo, seborrheic keratosis, and lichen planus–like keratosis), dermatofibroma, melanoma, invasive squamous cell carcinoma and keratoacanthoma, benign sebaceous neoplasms, and benign hair follicle tumors. The **Table** shows the frequencies of diagnoses in the training and validation set. The test set images of 2072 dermoscopic and clinical close-up images originated from multiple sources, including the Medical University of Vienna, the image database from

Table. Summary of Diagnoses in Training and Validation Data Sets

| Data Set | Total No. of Images | Images, No. (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AKIEC | Angioma | BCC | BKL | DF | Mel | Nevus | SCC | Seb-Ben | Trich-Ben |
| Training | | | | | | | | | | | |
| Dermoscopy | 7895 | 1892 (24.0) | 26 (0.3) | 3855 (48.8) | 891 (11.3) | 56 (0.7) | 58 (0.7) | 119 (1.5) | 957 (12.1) | 18 (0.2) | 23 (0.3) |
| Close-up | 5829 | 1379 (23.7) | 16 (0.3) | 2832 (48.6) | 668 (11.5) | 31 (0.5) | 37 (0.6) | 104 (1.8) | 762 (13.1) | 0 | 0 |
| Validation | | | | | | | | | | | |
| Dermoscopy | 340 | 8 (2.4) | 4 (1.2) | 165 (48.5) | 41 (12.1) | 6 (1.8) | 15 (4.4) | 7 (2.1) | 88 (25.9) | 3 (0.9) | 3 (0.9) |
| Close-up | 635 | 16 (2.5) | 7 (1.1) | 321 (50.6) | 80 (12.6) | 7 (1.1) | 25 (3.9) | 10 (1.6) | 169 (26.6) | 0 | 0 |

Abbreviations: AKIEC, actinic keratoses and intraepithelial carcinoma (also known as Bowen disease); BCC, basal cell carcinoma, all subtypes; BKL, benign keratosis–like lesions including solar lentigo, seborrheic keratosis, and lichen planus–like keratosis; DF, dermatofibroma; Mel, melanoma; SCC, invasive squamous cell carcinoma and keratoacanthoma; Seb-Ben, benign sebaceous neoplasms; Trich-Ben, benign hair follicle tumors.

C.R., and a convenience sample of rare diagnoses. The specific composition of the test set has already been described in greater detail by Sinz et al.[5] The list of diagnoses of the test set is available in eTable 1 in the Supplement.

### Neural Network Diagnoses

The output of our trained neural networks represents probability values between 0 and 1 for every diagnostic category. We combined the outputs of 2 CNNs (eAppendix in the Supplement), one trained with dermoscopic images and the other with clinical close-ups, by extreme gradient boosting (XGBoost[25]) of the combined probabilities. This combined model is referred to as cCNN. For specific diagnoses, we used the highest combined class probability and summed the probabilities of malignant and benign categories to generate receiver operating characteristic curves. In the validation set, we fixed the specificity at the level of the average human rater (51.3%), which corresponded to a combined malignant class probability of 0.2 (eFigure in the Supplement).

### Human Ratings

The specifics of the rater study have been described in detail by Sinz et al.[5] In a web-based study of 95 human raters (51.6% female; mean age, 43.4 years; 95% CI, 41.0-45.7 years), participants were divided into 3 groups (according to years of experience with dermoscopy): beginner raters (<3 years), intermediate raters (3-10 years), or expert raters (>10 years). Data on the formal education of the rates are shown in eTable 2 in the Supplement. All participants rated 50 cases drawn randomly from the entire test set of 2072 nonpigmented lesions. The random sample was stratified according to diagnostic category to prevent overrepresentation of common diagnoses. The raters were asked to differentiate between benign and malignant lesions, to make a specific diagnosis, and to suggest therapeutic management. The clinical close-up image was always shown before the dermatoscopic image, and the final evaluation was based on the combination of both imaging modalities.

### Statistical Analysis

Statistical calculations and visualizations were performed with R Statistics.[26] The 50 randomly drawn ratings of each rater were compared with the cCNN output for the same 50 cases in a pairwise fashion using paired *t* tests. Receiver operating characteristic curves were calculated by pooling all rating sets to allow comparability with results from the human raters. The receiver operating characteristic curves and the area under the curves (AUC) were calculated using pROC,[27] and a comparison of receiver operating characteristic curves was performed using the methods of DeLong et al.[28] The primary end point for the analyses was difference in AUC to detect skin cancer between the CNN and human raters. All reported *P* values were from 2-sided tests and are corrected for multiple testing with the Benjamini-Hochberg method[29] and are considered statistically significant at a corrected value of $P < .05$.

## Results

### CNN Training and Validation

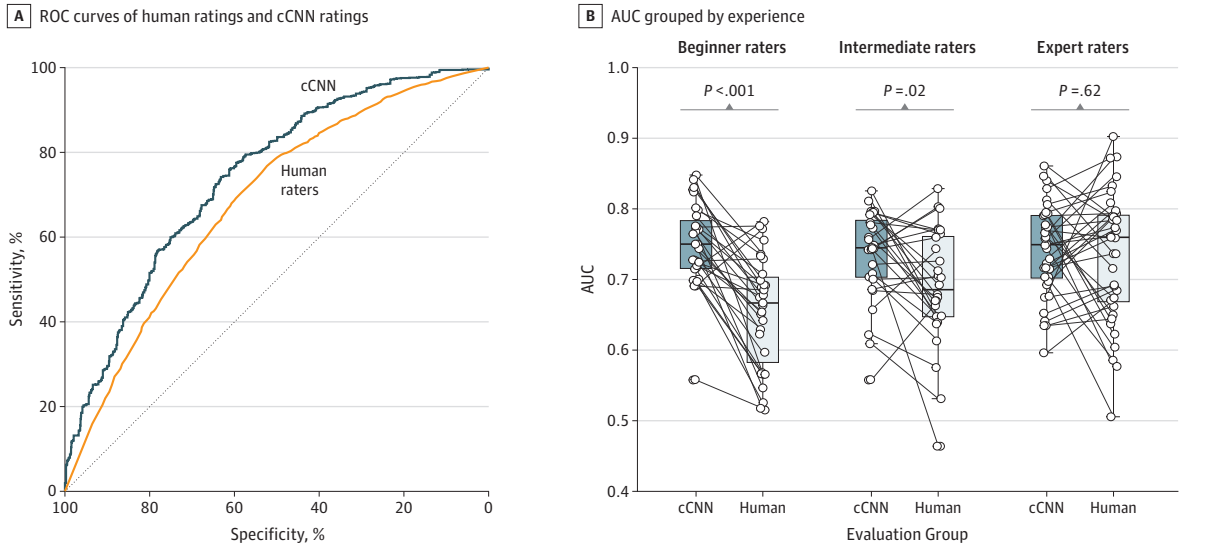
For the validation set, an InceptionV3 architecture[30] achieved the highest accuracy rates for dermoscopic images, and a ResNet50 network[31] had the highest accuracy for clinical close-ups. With regard to the detection of skin cancer, the AUC of the CNN was significantly higher with dermoscopy than with clinical close-ups (0.725; 95% CI, 0.711-0.725 vs 0.683; 95% CI, 0.668-0.683; $P < .001$). Regarding specific diagnoses, the dermoscopic CNN was better at diagnosing malignant cases, and the close-up CNN was better at diagnosing benign cases (eTable 3 in the Supplement). Integration of both methods using extreme gradient boosting achieved significantly higher accuracy than dermoscopy alone (AUC, 0.742; 95% CI, 0.729-0.755; $P < .001$). The rate of correct specific diagnoses was highest in the combined ratings (cCNN, 37.6% vs close-up CNN, 31.1%; $P < .001$; and dermoscopic CNN, 36.3%; $P = .005$).

### Comparison of Human Raters With cCNN

With regard to the detection of skin cancer, the mean AUC of human raters (0.695; 95% CI, 0.676-0.713) was significantly lower than the mean AUC of the cCNN (0.742; 95% CI, 0.729-0.755; $P < .001$) (**Figure 1**A). Comparing subgroups (Figure 1B), we found that the CNN had a higher AUC than did beginner raters (0.749; 95% CI, 0.727-0.771 vs 0.655; 95% CI, 0.626-0.684; $P < .001$) or the intermediate raters (0.735; 95% CI, 0.710-0.760; vs 0.690; 95% CI, 0.657-0.722; $P = .02$), but not higher than the experts (0.733; 95% CI, 0.702-0.765 vs 0.741; 95% CI, 0.719-0.763; $P = .62$).

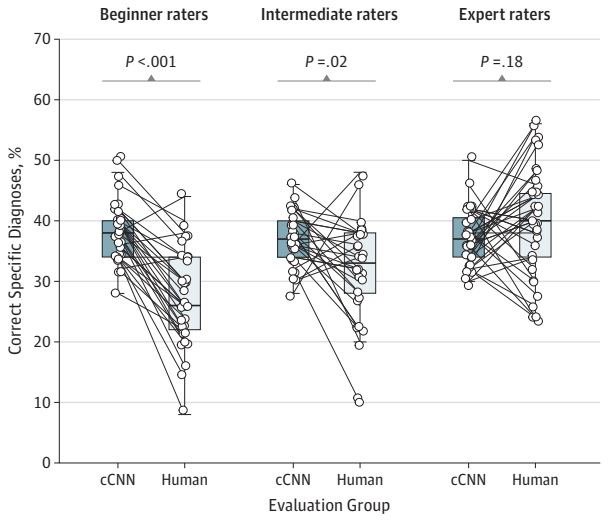
Although sensitivity was 77.6% (95% CI, 74.7%-80.5%) and specificity was 51.3% (95% CI, 48.4%-54.3%) for human raters,

Figure 1. Comparison of Skin Cancer Detection on Digital Images Between Human Readers and a Neural Network–Based Classifier




A, Receiver operating characteristic (ROC) curves of pooled human ratings (orange) and the combined convolutional neural network (cCNN) rating (blue) show significantly higher performance by the automated classifier. B, Area under the curve (AUC) of corresponding reading sets of the cCNN and dermatologists, grouped by experience. The horizontal line in each box indicates the median (middle band), while the top and bottom borders of the box indicate the 75th and 25th percentiles, respectively.

Figure 2. Percentages of Correct Specific Diagnoses of Corresponding Reading Sets of the Combined Convolutional Neural Network (cCNN) and Dermatologists Grouped by Experience




The horizontal line in each box indicates the median (middle band), while the top and bottom borders of the box indicate the 75th and 25th percentiles, respectively.

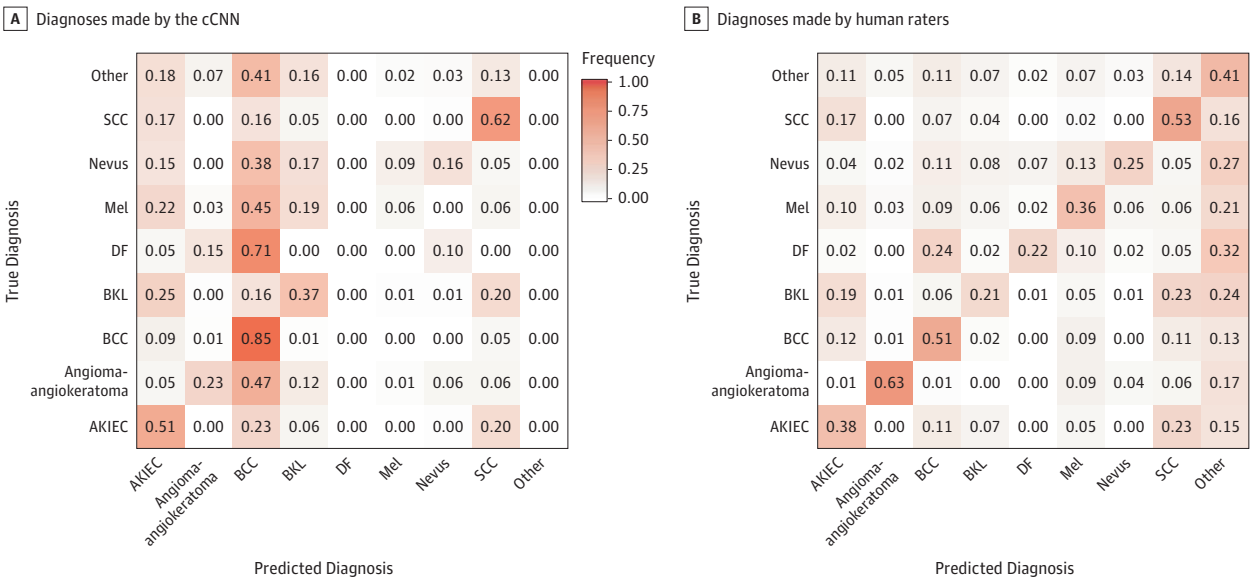
the values for the cCNN were higher but not significantly different (sensitivity, 80.5%; 95% CI, 79.0%-82.1%; $P = .12$; specificity, 53.5%; 95% CI, 51.7%-55.3%; $P = .298$). Except for a significantly higher sensitivity of the neural network compared with beginner raters (81.9%; 95% CI, 79.2%-84.6%; vs 72.3%; 95% CI, 66.7%-77.9%; $P = .003$), there were no significant differences with other subgroups of human raters. Regarding the rare, but important, class of primary amelanotic melanoma, the cCNN achieved a sensitivity of 52.3% (95% CI, 47.2%-57.4%) when diagnosing malignancy, which was lower than the sensitivity reached by beginner raters (59.8%; 95% CI, 50.6%-68.5%), intermediate raters (67.8%; 95% CI, 58.5%-75.9%), and expert raters (78.5%; 95% CI, 70.7%-84.7%).

The cCNN, combining analysis of dermoscopy images and clinical close-ups, achieved a higher frequency of correct specific diagnoses (37.6%; 95% CI, 36.6%-38.4%) than did human raters (33.5%; 95% CI, 31.5%-35.6%; $P = .001$). This difference was significant for beginner raters and intermediate raters but not expert raters (37.3%; 95% CI, 35.7%-38.8% vs 40.0%; 95% CI, 37.0%-43.0%; $P = .18$) (**Figure 2**). With regard to specific diagnoses, the difference between cCNN and human raters was higher when only malignant lesions were considered (55.5%; 95% CI, 54.0%-57.1% vs 44.9%; 95% CI, 42.2%-47.7%; $P < .001$). When the analysis was limited to benign cases, human raters were significantly better (frequency of correct specific diagnoses: 23.4%; 95% CI, 20.8%-25.9% vs 18.1%; 95% CI, 16.8%-19.3%; $P = .001$). Confusion matrices (**Figure 3**) demonstrate that the cCNN performs better on common malignant classes (actinic keratoses and intraepithelial carcinoma [Bowen disease], basal cell carcinoma, and invasive squamous cell carcinoma and keratoacanthoma) but performs poorly on benign classes such as angiomas, dermatofibromas, nevi, or clear cell acanthomas (**Figure 4**), which were underrepresented or absent in the training data.

## Discussion

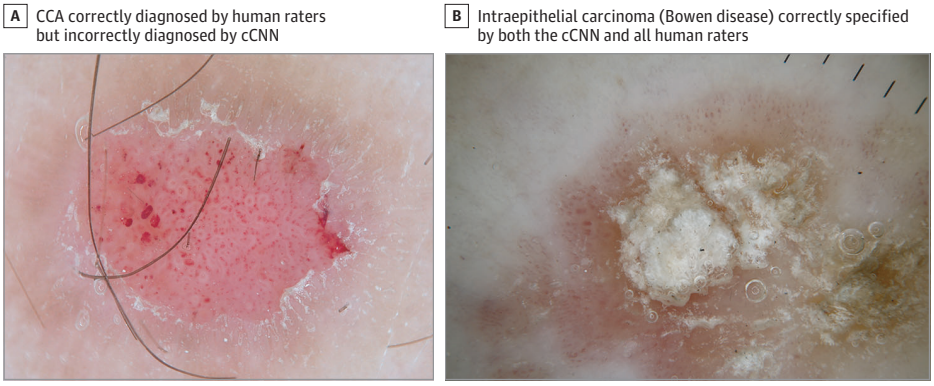
We showed that a cCNN is able to classify nonpigmented lesions as accurately as expert raters and with a higher accuracy than less-experienced raters. Because we used dermoscopic images and clinical close-ups to train the network, our results also dem-

Figure 3. Confusion Matrices of Specific Diagnoses




A, Diagnoses made by the combined convolutional neural network (cCNN). B, Diagnoses made by human raters. Values normalized to ground truth (rows). For reference, see frequency gradient scale used in panel A. AKIEC indicates actinic keratoses and intraepithelial carcinoma (also known as Bowen disease); BCC, basal cell carcinoma (all subtypes); BKL, benign keratosis-like lesions (including solar lentigo, seborrheic keratosis and lichen planus–like keratosis); DF, dermatofibroma; Mel, melanoma; and SCC, invasive squamous cell carcinoma.

Figure 4. Example Images




A, A clear cell acanthoma (CCA) correctly diagnosed by all human raters, but interpreted as a benign keratosis-like lesion by the combined convolutional neural network (cCNN). Since the class CCA was not present in the training data set it is impossible for the fixed classifier to ever make that diagnosis. B, An actinic keratosis and intraepithelial carcinoma (also known as Bowen disease) correctly specified by both the cCNN and all human raters.

onstrate that a combination of the 2 imaging modalities achieves better results than either modality alone. In this regard, we confirmed the importance of adding the dermoscopic images to the clinical examination and the importance of considering the clinical close-up images in addition to the dermoscopic images and not to rely on the dermoscopic images alone.[32] The 2 methods complement each other. The CNN analyzing close-up images was more accurate for benign lesions, whereas the CNN analyzing the corresponding dermoscopic images was more accurate for malignant cases (eTable 3 in the Supplement).

Our experimental setting was artificial and deviated from clinical practice in many ways. It was restricted to pure morphologic characteristics, and we did not include important metadata such as age, anatomic site, and history of the lesions. These data will usually be readily available to the treating physician and will affect diagnosis and management. In this regard, we see the strength of CNN-based classifiers not so much in providing management decisions[33] but rather in providing a list of accurate differential diagnoses, which may serve as input for other systems that have outputs, such as decision trees, that are more readily interpretable by humans.

Our data also suffer from verification bias, as only pathologically verified cases were selected. This selection leads to overrepresentation of malignant cases and an unequal class distribution in the test set, which does not reflect clinical reality. Dermatopathologic verification, however, is necessary because the clinical and dermoscopic diagnosis of nonpigmented lesions is prone to error, and we think that the advantage of an accurate criterion-standard diagnosis outweighs the disadvantage of verification bias.

The performance of the cCNN was not uniform across classes. It outperformed human raters in common malignant classes such as basal cell carcinoma, actinic keratoses or Bowen disease, and squamous cell carcinoma or keratoacanthoma but did not reach the accuracy of human raters in rare malignant nonpigmented lesions such as amelanotic melanoma and benign nonpigmented lesions. This is a consequence of the relatively low frequency of these disease categories in the training set. Although this, to our knowledge, is the largest data set of nonpigmented dermoscopic images, it still counts as a small data set in the realm of machine learning with CNNs. During a professional life, a typical human expert rater has been exposed to a significant number of exemplars, even for rare diagnoses, either through textbooks, e-learning, lectures, or clinical practice.

The rare but important class of amelanotic melanoma is difficult to diagnose even for experts. Usually no harm is done if amelanotic melanomas are mistaken for other malignant neoplasms or if, in the judgement of the physician, the probability of a malignant neoplasm is high enough to warrant biopsy or excision. Dermatoscopy is more accurate when classifying amelanotic melanoma as malignant rather than melanoma.[34] Assuming that, if the diagnosis of the cCNN is a malignant neoplasm, the lesion will be biopsied or excised, the cCNN achieved a sensitivity for amelanotic melanoma of 52.3% (95% CI, 47.2%-57.4%), which was lower than the average sensitivity of human raters of 69.3% (95% CI, 64.5%-73.8%). We hypothesize that, in addition to underrepresentation of amelanotic melanomas in the training data, visual diagnostic clues such as polymorphous vessels are too subtle to be learned from just a few cases. Unless larger image collections become available, other diagnostic devices such as reflectance confocal microscopy or automated diagnostic systems that do not depend on morphologic characteristics (eg, tapestripping,[35] electrical impedance spectroscopy,[36] or Raman spectroscopy)[37] may be of more help in these cases.

### Limitations

The lower accuracy of the presented cCNN compared with other recent publications on automated classification of skin lesions[16,18] may be explained in 2 ways. One is that our test set included more than 51 distinct classes, of which most did not have enough examples to be integrated into the training phase. Having larger dermoscopy data sets in the future, in the scale of the number of clinical images that were available to Han et al,[14] may partly resolve this shortcoming. Second, the features of nonpigmented lesions are less specific than those of pigmented lesions, which is mirrored by the relatively low accuracy of human expert raters. Although our cCNN outperformed human raters in some aspects, it is currently not fit for clinical application. The metrics applied to measure diagnostic accuracy, such as sensitivity, specificity, and area under receiver operating characteristic curves, may not accurately reflect the performance of a classifier for medical purposes in all settings. Accurate diagnoses of common diseases such as basal cell carcinoma and actinic keratoses, which are usually not life threatening if left untreated, must be contrasted with missing potentially life-threatening diseases such as amelanotic melanomas. Although metrics exist that take into account the potential loss of life-years and apply penalties to misdiagnoses of more aggressive diseases, these metrics are currently not well established in the field of machine learning.

## Conclusions

Despite limitations, we demonstrated that CNNs can perform at a human level on the binary classification of pigmented lesions and on multiclass tasks on more challenging nonpigmented lesions. The potential of CNNs to solve more sophisticated classification tasks in dermatology has been demonstrated before[13,14] but not on dermoscopic images of nonpigmented lesions. We also confirm that, similar to human raters, CNNs perform better with dermoscopic images than with clinical close-ups alone. Future efforts should be targeted at the availability of larger numbers of dermoscopic images and clinical close-ups of rare malignant lesions but also of common benign nonpigmented lesions that are usually not biopsied or excised for diagnostic reasons. The results of our study suggest that, if more exemplars of these disease categories were available, it should be possible to train CNNs to diagnose these categories more efficiently.

ARTICLE INFORMATION

**Accepted for Publication:** September 21, 2018.

**Published Online:** November 28, 2018. doi:10.1001/jamadermatol.2018.4378

**Author Affiliations:** School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada (Tschandl); Vienna Dermatologic Imaging Research Group, Department of Dermatology, Medical University of Vienna, Vienna, Austria (Tschandl, Neuber, Sinz, Kittler); School of Medicine, The University of Queensland, Brisbane, Queensland, Australia (Rosendahl); School of Medicine, Tehran University of Medical Sciences, Tehran, Iran (Rosendahl); Department of Dermatology, Ankara University Faculty of Medicine, Ankara, Turkey (Akay); Dermatology Unit, University of Campania, Naples, Italy (Argenziano); Public, Private and Teaching Practice of Dermatology, Konstanz, Germany (Blum); Department of Dermatology, University Hospital Zürich, Zürich, Switzerland (Braun); Department of Dermatology, Instituto de Investigaciones Médicas ALanari, University of Buenos Aires, Buenos Aires, Argentina (Cabo); Centre de Dermatologie, Nemours, France (Gourhant); private practice, Lübeck, Germany (Kreusch); First Department of Dermatology, Aristotle University, Thessaloniki, Greece (Lallas); Department of Dermatology, Karolinska University Hospital and Karolinska Institutet, Stockholm, Sweden (Lapins); Dermatology Service, Memorial Sloan Kettering Cancer Center, Hauppauge, New York (Marghoob); Sydney Melanoma Diagnostic Centre and Discipline of Dermatology, University of Sydney, Sydney, Australia (Menzies); Department of Dermatology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden (Paoli); Skin and Cancer Associates, Plantation, Florida (Rabinovitz); Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria (Rinner); Medical Screening Institute, Chaim Sheba Medical Center, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel (Scope); Dermatology Research Centre, The University of Queensland, The University of Queensland Diamantina Institute, Brisbane, Australia (Soyer); Department of Dermatology, Centre Hospitalier Lyon Sud, Lyon 1 University, Lyons Cancer Research Center, Lyon, France (Thomas); Dermatology Clinic, Maggiore Hospital, University of Trieste, Trieste, Italy (Zalaudek).

**Author Contributions:** Drs Tschandl and Kittler had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
*Concept and design:* Tschandl, Neuber, Soyer, Kittler.
*Acquisition, analysis, or interpretation of data:* Tschandl, Rosendahl, Akay, Argenziano, Blum, Braun, Cabo, Gourhant, Kreusch, Lallas, Lapins, Marghoob, Menzies, Neuber, Paoli, Rabinoviz, Rinner, Scope, Sinz, Thomas, Zalaudek, Kittler.

*Drafting of the manuscript:* Tschandl, Kittler.
*Critical revision of the manuscript for important intellectual content:* All authors.
*Statistical analysis:* Tschandl.
*Obtained funding:* Kittler.
*Administrative, technical, or material support:* Akay, Kreusch, Lapins, Neuber, Rinner, Zalaudek, Kittler.
*Supervision:* Cabo, Paoli, Kittler.

**Conflict of Interest Disclosures:** Dr Tschandl reported receiving an unrestricted grant from MetaOptima Technology Inc for conducting a 1-year postdoctoral fellowship at Simon Fraser University, Burnaby, British Columbia, Canada.

**Additional Contributions:** We thank all the raters who participated in online assessment of skin lesions, without whom this study would not have been possible.

## REFERENCES


1. Kittler H, Marghoob AA, Argenziano G, et al. Standardization of terminology in dermoscopy/dermatoscopy: results of the third consensus conference of the International Society of Dermoscopy. *J Am Acad Dermatol*. 2016;74(6):1093-1106. doi:10.1016/j.jaad.2015.12.038

2. Argenziano G, Cerroni L, Zalaudek I, et al. Accuracy in melanoma detection: a 10-year multicenter survey. *J Am Acad Dermatol*. 2012;67(1):54-59. doi:10.1016/j.jaad.2011.07.019

3. Rosendahl C, Cameron A, Bulinska A, Williamson R, Kittler H. Dermatoscopy of a minute melanoma. *Australas J Dermatol*. 2011;52(1):76-78. doi:10.1111/j.1440-0960.2010.00725.x

4. Rosendahl C, Tschandl P, Cameron A, Kittler H. Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. *J Am Acad Dermatol*. 2011;64(6):1068-1073. doi:10.1016/j.jaad.2010.03.039

5. Sinz C, Tschandl P, Rosendahl C, et al. Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. *J Am Acad Dermatol*. 2017;77(6):1100-1109. doi:10.1016/j.jaad.2017.07.022

6. Binder M, Steiner A, Schwarz M, Knollmayer S, Wolff K, Pehamberger H. Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. *Br J Dermatol*. 1994;130(4):460-465. doi:10.1111/j.1365-2133.1994.tb03378.x

7. Rubegni P, Cevenini G, Burroni M, et al. Automated diagnosis of pigmented skin lesions. *Int J Cancer*. 2002;101(6):576-580. doi:10.1002/ijc.10620

8. Menzies SW, Bischof L, Talbot H, et al. The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma [published correction appears in *Arch Dermatol*. 2006;142(5):558]. *Arch Dermatol*. 2005;141(11):1388-1396. doi:10.1001/archderm.141.11.1388

9. Dreiseitl S, Binder M, Vinterbo S, Kittler H. Applying a decision support system in clinical practice: results from melanoma diagnosis. *AMIA Annu Symp Proc*. 2007;191-195.

10. Binder M, Kittler H, Dreiseitl S, Ganster H, Wolff K, Pehamberger H. Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process. *Melanoma Res*. 2000;10(6):556-561. doi:10.1097/00008390-200012000-00007

11. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*. Vol. 25. Red Hook, NY: Curran Associates Inc; 2012:1097-1105.

12. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115(3):211-252. doi:10.1007/s11263-015-0816-y

13. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

14. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol*. 2018;138(7):1529-1538. doi:10.1016/j.jid.2018.01.028

15. Tschandl P, Kittler H, Argenziano G. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. *Br J Dermatol*. 2017;177(3):867-869. doi:10.1111/bjd.15695

16. Marchetti MA, Codella NCF, Dusza SW, et al; International Skin Imaging Collaboration. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol*. 2018;78(2):270-277.e1. doi:10.1016/j.jaad.2017.08.016

17. Yu C, Yang S, Kim W, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. *PLoS One*. 2018;13(3):e0193321. doi:10.1371/journal.pone.0193321

18. Haenssle HA, Fink C, Schneiderbauer R, et al; Reader study level-I and level-II Groups. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29(8):1836-1842. doi:10.1093/annonc/mdy166

19. Ge Z, Demyanov S, Chakravorty R, Bowling A, Garnavi R. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In: Descoteaux M, Maier-Hein L, Franz J, et al, eds. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2017*. Cham, Switzerland: Springer International Publishing; 2017:250-258. doi:10.1007/978-3-319-66179-7_29

20. Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. *Exp Dermatol*. 2018;27(11):1261-1267. doi:10.1111/exd.13777

21. Kharazmi P, Kalia S, Lui H, Wang ZJ, Lee TK. A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile. *Skin Res Technol*. 2018;24(2):256-264. doi:10.1111/srt.12422

22. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. 7-Point checklist and skin lesion classification using multi-task multi-modal neural nets [published online April 9, 2018]. *IEEE J Biomed Health Inform*. doi:10.1109/JBHI.2018.2824327

23. Ballerini L, Fisher RB, Aldridge B, Rees J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Celebi ME, Schaefer G, eds. *Color Medical Image Analysis*. Dordrecht, the Netherlands: Springer; 2013:63-86. doi:10.1007/978-94-007-5389-1_4

24. Finnane A, Curiel-Lewandrowski C, Wimberley G, et al; International Society of Digital Imaging of the Skin (ISDIS) for the International Skin Imaging Collaboration (ISIC). Proposed technical guidelines for the acquisition of clinical images of skin-related conditions. *JAMA Dermatol*. 2017;153(5):453-457. doi:10.1001/jamadermatol.2016.6214

25. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. arXiv [csLG] 2016. https://arxiv.org/abs/1603.02754. Accessed October 17, 2018.

26. R Core Team. R: a language and environment for statistical computing. https://www.R-project.org/. 2017. Accessed October 17, 2018.

27. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. doi:10.1186/1471-2105-12-77

28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. doi:10.2307/2531595

29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289-300.

30. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception architecture for computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, Nevada:2818-2826.

31. He K, Xiangyu Z, Shaoqing R, Jian S. Deep residual learning for image recognition. arXiv [Cs] 2015. http://arxiv.org/abs/1512.03385. Accessed October 17, 2018.

32. Carli P, de Giorgi V, Chiarugi A, et al. Addition of dermoscopy to conventional naked-eye examination in melanoma screening: a randomized study. *J Am Acad Dermatol*. 2004;50(5):683-689. doi:10.1016/j.jaad.2003.09.009

33. Cook DA, Sherbino J, Durning SJ. Management reasoning: beyond the diagnosis. *JAMA*. 2018;319(22):2267-2268. doi:10.1001/jama.2018.4385

34. Menzies SW, Kreusch J, Byth K, et al. Dermoscopic evaluation of amelanotic and hypomelanotic melanoma. *Arch Dermatol*. 2008;144(9):1120-1127. doi:10.1001/archderm.144.9.1120

35. Wachsman W, Morhenn V, Palmer T, et al. Noninvasive genomic detection of melanoma. *Br J Dermatol*. 2011;164(4):797-806. doi:10.1111/j.1365-2133.2011.10239.x

36. Malvehy J, Hauschild A, Curiel-Lewandrowski C, et al. Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety. *Br J Dermatol*. 2014;171(5):1099-1107. doi:10.1111/bjd.13121

37. Lui H, Zhao J, McLean D, Zeng H. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. *Cancer Res*. 2012;72(10):2491-2500. doi:10.1158/0008-5472.CAN-11-4061